

Skew- t Filter and Smoother with Improved Covariance Matrix Approximation

Henri Nurminen, Tohid Ardeshtiri, Robert Piché, and Fredrik Gustafsson

Abstract—Filtering and smoothing algorithms for linear discrete-time state-space models with skew- t -distributed measurement noise are presented. The presented algorithms use a variational Bayes based posterior approximation with coupled location and skewness variables to reduce the error caused by the variational approximation. Although the variational update is done suboptimally, our simulations show that the proposed method gives a more accurate approximation of the posterior covariance matrix than an earlier proposed variational algorithm. Consequently, the novel filter and smoother outperform the earlier proposed robust filter and smoother and other existing low-complexity alternatives in accuracy and speed. We present both simulations and tests based on real-world navigation data, in particular GPS data in an urban area, to demonstrate the performance of the novel methods. Moreover, the extension of the proposed algorithms to cover the case where the distribution of the measurement noise is multivariate skew- t is outlined. Finally, the paper presents a study of theoretical performance bounds for the proposed algorithms.

Index Terms— skew t , skewness, t -distribution, robust filtering, Kalman filter, variational Bayes, RTS smoother, truncated normal distribution, Cramér–Rao lower bound

I. INTRODUCTION

Asymmetric and heavy-tailed noise processes are present in many inference problems. In radio signal based distance estimation [1]–[3], for example, obstacles cause large positive errors that dominate over symmetrically distributed errors from other sources [4]. An example of this is the error histogram of time-of-flight in distance measurements collected in an indoor environment given in Fig. 1. The asymmetric outlier distributions cannot be predicted by the normal distribution that is equivalent in second order moments, because normal distributions are symmetric thin-tailed distributions. The skew t -distribution [5]–[7] is a generalization of the t -distribution that has the modeling flexibility to capture both skewness and heavy-tailedness of such noise processes. To illustrate this, Fig. 2 shows the contours of the likelihood function for three range measurements where some of the measurements are positive outliers. In this example, skew- t , t , and normal measurement noise models are compared. Due to the additional modeling flexibility, the skew- t based likelihood provides a more apposite spread of the probability mass than the normal and t based likelihoods.

H. Nurminen and R. Piché are with the Department of Automation Science and Engineering, Tampere University of Technology (TUT), PO Box 692, 33101 Tampere, Finland (e-mails: henri.nurminen@tut.fi, robert.piche@tut.fi). H. Nurminen receives funding from TUT Graduate School, the Foundation of Nokia Corporation, and Tekniikan edistämisyhdistys.

T. Ardeshtiri was with the Division of Automatic Control, Department of Electrical Engineering, Linköping University, 58183, Linköping, Sweden and received funding from Swedish research council (VR), project scalable Kalman filters for this work. T. Ardeshtiri is currently with the Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK, (e-mail: ta417@cam.ac.uk).

F. Gustafsson is with the Division of Automatic Control, Department of Electrical Engineering, Linköping University, 58183 Linköping, Sweden, (e-mail: fredrik@isy.liu.se).

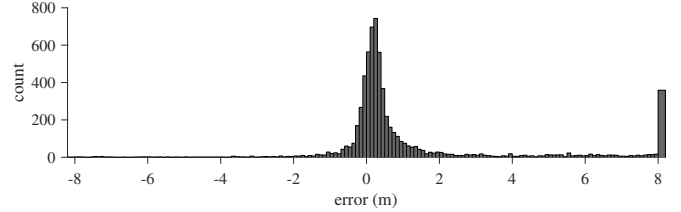


Fig. 1. The error histogram in an ultra-wideband (UWB) ranging experiment described in [8] shows positive skewness. The edge bars show the errors outside the figure limits.

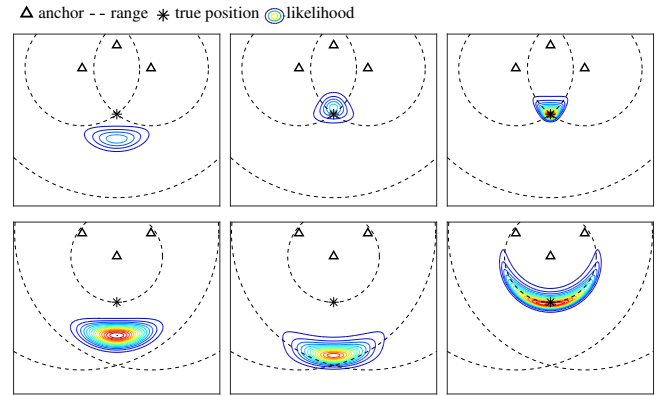


Fig. 2. The contours of the likelihood function for three range measurements for the normal (left), t (middle) and skew- t (right) measurement noise models. The t and skew- t based likelihoods handle one outlier (upper row), while only the skew- t model handles the two positive outlier measurements (bottom row) due to its asymmetry. The measurement model parameters are selected such that the degrees-of-freedom values and the first two moments coincide.

The applications of the skew distributions are not limited to radio signal based localization. In biostatistics skewed distributions are used as a modeling tool for handling heterogeneous data involving asymmetric behaviors across subpopulations [9]. In psychiatric research skew normal distribution is used to model asymmetric data [10]. Further, in economics skew normal and skew t -distributions are used as models for describing claims in property-liability insurance [11]. More examples describing approaches for analysis and modeling using multivariate skew normal and skew t -distributions in econometrics and environmetrics are presented in [12].

There are various algorithms dedicated to statistical inference of time series when the data exhibit asymmetric distribution. Particle filters [13] can easily be adapted to skew noise distributions, but the computational complexity of these filters increases rapidly as the state dimension increases. A skew Kalman filter is proposed in [14], and in [15] this filter is extended to a robust scale-mixture filter using Monte Carlo integration. These solutions are based on state-space models where the measurement noise is a dependent process with skewed marginals. The article [16] proposes filtering of independent skew measurement and process noises with the

cost of increasing the filter state's dimension over time. In all the skew filters of [14]–[16], sequential processing requires numerical evaluation of multidimensional integrals. The inference problem with skew likelihood distributions can also be cast into an optimization problem; [3] proposes an approach to model the measurement noise in an ultra-wideband (UWB) based positioning problem using a tailored half-normal-half-Cauchy distribution. Skewness can also be modeled by a mixture of normal distributions (Gaussian mixtures, GM) [1]. There are many filtering algorithms for GM distributions such as Gaussian sum filter [17] and interactive multiple model (IMM) filter [18]. However, GMs have exponentially decaying tails and can thus be too sensitive to outlier measurements. Furthermore, in order to keep the computational cost of a Gaussian sum filter practicable, a mixture reduction algorithm (MRA) [19] is required, and these MRAs can be computationally expensive and involve approximations to the posterior density. Filtering and smoothing algorithms for linear discrete-time state-space models with skew- t measurement noise using a variational Bayes (VB) method are presented in [20]. In tests with real UWB indoor localization data [8], this filter was shown to be accurate and computationally inexpensive.

This paper proposes improvements to the robust filter and smoother proposed in [20]. Analogous to [20], the measurement noise is modeled by the skew t -distribution, and the proposed filter and smoother use a VB approximation of the filtering and smoothing posteriors. However, the main contributions of this paper are (1) a new factorization of the approximate posterior distribution, (2) derivation of Cramér–Rao lower bound (CRLB) for the proposed filter and smoother, (3) the application of an existing method for approximating the statistics of a truncated multivariate normal distribution (TMND), and (4) a proof of optimality for a truncation ordering in approximation of the moments of the TMND. A TMND is a multivariate normal distribution whose support is restricted (truncated) by linear constraints and that is re-normalized to integrate to unity. The aforementioned contributions improve the estimation performance of the skew- t filter and smoother by reducing the covariance underestimation common to most VB inference algorithms [21, Chapter 10]. To our knowledge, VB approximations have been applied to the skew t -distribution only in our earlier works [8], [20] and by Wand et al. [22].

The rest of this paper is structured as follows. In Section II, the filtering and smoothing problem involving the univariate skew t -distribution is posed. In Section III a solution based on VB for the formulated problem is proposed. The proposed solution is evaluated using simulated data as well as real-world data in Sections IV and V, respectively. The essential expressions to extend the proposed filtering and smoothing algorithms to problems involving multivariate skew- t (MVST) distribution are given in Section VI. A performance bound for time series data with MVST-distributed measurement noise is derived and evaluated in simulation in Section VII. The concluding remarks are given in Section VIII.

II. INFERENCE PROBLEM FORMULATION

Consider the linear and Gaussian state evolution model

$$p(x_1) = \mathcal{N}(x_1; x_{1|0}, P_{1|0}), \quad (1a)$$

$$x_{k+1} = Ax_k + w_k, \quad w_k \stackrel{\text{iid}}{\sim} \mathcal{N}(0, Q), \quad (1b)$$

where $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes the probability density function (PDF) of the (multivariate) normal distribution with mean μ

and covariance matrix Σ ; $A \in \mathbb{R}^{n_x \times n_x}$ is the state transition matrix; $x_k \in \mathbb{R}^{n_x}$ indexed by $1 \leq k \leq K$ is the state to be estimated with initial prior distribution (1a), where the subscript “ $a|b$ ” is read “at time a using measurements up to time b ”; and $w_k \in \mathbb{R}^{n_x}$ is the process noise. Further, the measurements $y_k \in \mathbb{R}^{n_y}$ are assumed to be governed by the measurement equation

$$y_k = Cx_k + e_k, \quad (2)$$

where $C \in \mathbb{R}^{n_y \times n_x}$ is the measurement matrix, and the measurement noise e_k is independent of the process noise and has the product of independent univariate skew t -distributions as the PDF

$$[e_k]_i \stackrel{\text{iid}}{\sim} \text{ST}(0, R_{ii}, \Delta_{ii}, \nu_i). \quad (3)$$

The model can also be nonstationary, but for the sake of lighter notation the k subscripts on A , Q , C , R_{ii} , Δ_{ii} , and ν_i are omitted. The univariate skew t -distribution $\text{ST}(\mu, \sigma^2, \delta, \nu)$ is parametrized by its location parameter $\mu \in \mathbb{R}$, spread parameter $\sigma > 0$, shape parameter $\delta \in \mathbb{R}$ and degrees of freedom $\nu > 0$, and has the PDF

$$\text{ST}(z; \mu, \sigma^2, \delta, \nu) = 2 \mathfrak{t}(z; \mu, \sigma^2 + \delta^2, \nu) T(\tilde{z}; 0, 1, \nu + 1), \quad (4)$$

where

$$\mathfrak{t}(z; \mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma \sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{(z - \mu)^2}{\nu\sigma^2} \right)^{-\frac{\nu+1}{2}} \quad (5)$$

is the PDF of Student's t -distribution, $\Gamma(\cdot)$ is the gamma function, and $\tilde{z} = \frac{(z - \mu)\delta}{\sigma} \left(\frac{\nu+1}{\nu(\sigma^2 + \delta^2) + (z - \mu)^2} \right)^{\frac{1}{2}}$. Also, $T(\cdot; 0, 1, \nu)$ denotes the cumulative distribution function (CDF) of Student's t -distribution with degrees of freedom ν . Expressions for the first two moments of the univariate skew t -distribution can be found in [23].

The model (3) with independent univariate skew- t -distributed measurement noise components is justified in applications where one-dimensional data from different sensors can be assumed to have statistically independent noise [20]. Extension and comparison to multivariate skew- t -distributed noise will be discussed in Section VI.

The independent univariate skew- t noise model (3) induces the hierarchical representation of the measurement likelihood

$$y_k | x_k, u_k, \Lambda_k \sim \mathcal{N}(Cx_k + \Delta u_k, \Lambda_k^{-1} R), \quad (6a)$$

$$u_k | \Lambda_k \sim \mathcal{N}_+(0, \Lambda_k^{-1}), \quad (6b)$$

$$[\Lambda_k]_{ii} \sim \mathcal{G}(\frac{\nu_i}{2}, \frac{\nu_i}{2}), \quad (6c)$$

where $R \in \mathbb{R}^{n_y \times n_y}$ is a diagonal matrix whose diagonal elements' square roots $\sqrt{R_{ii}}$ are the spread parameters of the skew t -distribution in (3); $\Delta \in \mathbb{R}^{n_y \times n_y}$ is a diagonal matrix whose diagonal elements Δ_{ii} are the shape parameters; $\nu \in (\mathbb{R}^+)^{n_y}$ is a vector whose elements ν_i are the degrees of freedom; the operator $[\cdot]_{ij}$ gives the (i, j) entry of its argument; Λ_k is a diagonal matrix with a priori independent random diagonal elements $[\Lambda_k]_{ii}$. Also, $\mathcal{N}_+(\mu, \Sigma)$ is the TMND with closed positive orthant as support, location parameter μ , and squared-scale matrix Σ . Furthermore, $\mathcal{G}(\alpha, \beta)$ is the gamma distribution with shape parameter α and rate parameter β .

Bayesian smoothing means finding the smoothing posterior $p(x_{1:K}, u_{1:K}, \Lambda_{1:K} | y_{1:K})$. In [20], the smoothing posterior is approximated by a factorized distribution of the form

$q_{[20]} \triangleq q_x(x_{1:K})q_u(u_{1:K})q_\Lambda(\Lambda_{1:K})$. Subsequently, the approximate posterior distributions are computed using the VB approach. The VB approach minimizes the Kullback–Leibler divergence (KLD) $D_{\text{KL}}(q||p) \triangleq \int q(x) \log \frac{q(x)}{p(x)} dx$ [24] of the true posterior from the factorized approximation. That is, $D_{\text{KL}}(q_{[20]}||p(x_{1:K}, u_{1:K}, \Lambda_{1:K}|y_{1:K}))$ is minimized in [20].

The numerical simulations in [20] manifest covariance matrix underestimation, which is a known weakness of the VB approach [21, Chapter 10]. One of the contributions of this paper is to reduce the covariance underestimation of the filter and smoother proposed in [20] by removing independence approximations of the posterior approximation. The proposed filter and smoother are presented in Section III.

III. PROPOSED FILTER AND SMOOTHER

Using Bayes' theorem, the state evolution model (1), and the likelihood (6), the joint smoothing posterior PDF can be derived as in [20]. This posterior is not analytically tractable. We propose to seek an approximation in the form

$$p(x_{1:K}, u_{1:K}, \Lambda_{1:K}|y_{1:K}) \approx q_{xu}(x_{1:K}, u_{1:K}) q_\Lambda(\Lambda_{1:K}), \quad (7)$$

where the factors in (7) are specified by

$$\hat{q}_{xu}, \hat{q}_\Lambda = \underset{q_{xu}, q_\Lambda}{\operatorname{argmin}} D_{\text{KL}}(q_N||p(x_{1:K}, u_{1:K}, \Lambda_{1:K}|y_{1:K})) \quad (8)$$

and where $q_N \triangleq q_{xu}(x_{1:K}, u_{1:K})q_\Lambda(\Lambda_{1:K})$. Hence, $x_{1:K}$ and $u_{1:K}$ are not approximated as independent as in [20] because they can be highly correlated *a posteriori* [20]. The analytical solutions for \hat{q}_{xu} and \hat{q}_Λ are obtained by cyclic iteration of

$$\log q_{xu}(\cdot) \leftarrow \underset{q_\Lambda}{\mathbb{E}} [\log p(y_{1:K}, x_{1:K}, u_{1:K}, \Lambda_{1:K})] + c_{xu} \quad (9a)$$

$$\log q_\Lambda(\cdot) \leftarrow \underset{q_{xu}}{\mathbb{E}} [\log p(y_{1:K}, x_{1:K}, u_{1:K}, \Lambda_{1:K})] + c_\Lambda \quad (9b)$$

where the expected values on the right hand sides are taken with respect to the current q_{xu} and q_Λ [21, Chapter 10] [25], [26]. Also, c_{xu} and c_Λ are constants with respect to the variables $(x_{1:K}, u_{1:K})$ and $\Lambda_{1:K}$, respectively. Furthermore, the joint PDF $p(x_{1:K}, u_{1:K}, \Lambda_{1:K}, y_{1:K})$ can be written as

$$p(x_{1:K}, u_{1:K}, \Lambda_{1:K}, y_{1:K}) = p(x_1) \prod_{l=1}^{K-1} p(x_{l+1}|x_l) \\ \times \prod_{k=1}^K p(y_k|x_k, u_k, \Lambda_k) p(u_k|\Lambda_k) p(\Lambda_k) \quad (10)$$

$$= \mathcal{N}(x_1; x_{1|0}, P_{1|0}) \prod_{l=1}^{K-1} \mathcal{N}(x_{l+1}; Ax_l, Q) \\ \times \prod_{k=1}^K \mathcal{N}(y_k; Cx_k + \Delta u_k, \Lambda_k^{-1} R) \mathcal{N}_+(u_k; 0, \Lambda_k^{-1}) \\ \times \prod_{k=1}^K \prod_{i=1}^{n_y} \mathcal{G}([\Lambda_k]_{ii}; \frac{\nu_i}{2}, \frac{\nu_i}{2}). \quad (11)$$

Computation of the expectation in (9b), which is relegated to Appendix A, requires the first two moments of a TMND because the support of $u_{1:K}$ is the non-negative orthant. These moments can be computed using the formulas presented in [27]. They require evaluating the CDF of general multivariate normal distributions. The MATLAB function `mvncdf` implements the numerical quadrature of [28] in 2 and 3 dimensional cases and the quasi-Monte Carlo method of [29]

for the dimensionalities 4–25. However, these methods can be prohibitively slow. Therefore, we approximate the TMND's moments using the fast sequential algorithm suggested in [30], [31]. The method is initialized with the original normal density whose parameters are then updated by applying one linear constraint at a time. For each constraint, the mean and covariance matrix of the once-truncated normal distribution are computed analytically, and the once-truncated distribution is approximated by a non-truncated normal with the updated moments. This method is illustrated in Fig. 3, where a bivariate normal distribution truncated into the positive quadrant is approximated with a non-truncated normal distribution.

The result of the sequential truncation depends on the order in which the constraints are applied. Finding the optimal order of applying the truncations is a problem that has combinatorial complexity. Hence, we adopt a greedy approach, whereby the constraint to be applied is chosen from among the remaining constraints so that the resulting once-truncated normal is closest to the true TMND. By Lemma 1, the optimal constraint to select is the one that truncates the most probability. The optimality is with respect to a KLD as the measure. For example, in Fig. 3 the vertical constraint truncates more probability, so it is applied first.

Lemma 1. *Let $p(\mathbf{z})$ be a TMND with the support $\{\mathbf{z} \geq 0\}$ and $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mu, \Sigma)$. Then,*

$$\underset{i}{\operatorname{argmin}} D_{\text{KL}}\left(p(\mathbf{z}) \parallel \frac{1}{c_i} q(\mathbf{z}) \mathbb{I}[\mathbf{z}_i \geq 0]\right) = \underset{i}{\operatorname{argmin}} \frac{\mu_i}{\sqrt{\Sigma_{ii}}}, \quad (12)$$

where μ_i is the i th element of μ , Σ_{ii} is the i th diagonal element of Σ , $\mathbb{I}[\cdot]$ is the Iverson bracket, and $c_i = \int q(\mathbf{z}) \mathbb{I}[\mathbf{z}_i \geq 0] d\mathbf{z}$.

$$\text{Proof: } D_{\text{KL}}\left(p(\mathbf{z}) \parallel \frac{1}{c_i} q(\mathbf{z}) \mathbb{I}[\mathbf{z}_i \geq 0]\right) \\ \stackrel{\pm}{=} - \int p(\mathbf{z}) \log\left(\frac{1}{c_i} q(\mathbf{z}) \mathbb{I}[\mathbf{z}_i \geq 0]\right) d\mathbf{z} \quad (13) \\ = \log c_i - \int p(\mathbf{z}) \log q(\mathbf{z}) d\mathbf{z} - 1 \stackrel{\pm}{=} \log c_i, \quad (14)$$

where $\stackrel{\pm}{=}$ means equality up to an additive constant. Since c_i is an increasing function of $\frac{\mu_i}{\sqrt{\Sigma_{ii}}}$ the proof follows. ■

The obtained algorithm with the optimal processing sequence for computing the mean and covariance matrix of a given multivariate normal distribution truncated to the positive orthant is given in Table I. In many programming languages a numerically robust method to implement the line 6 of the algorithm in Table I is using the scaled complementary error function `erfcx` through

$$\frac{\phi(\xi)}{\Phi(\xi)} = \frac{\sqrt{2/\pi}}{\operatorname{erfcx}(-\xi/\sqrt{2})}. \quad (15)$$

The recursion (9) is convergent to a local optimum [21, Chapter 10]. However, there is no proof of convergence available when the moments of the TMND are approximated. In spite of lack of a convergence proof the iterations did not diverge in the numerical simulations presented in section IV.

In the smoother, the update (9a) includes a forward filtering step of the Rauch–Tung–Striebel smoother (RTSS) [32] to compute an approximate filtering posterior for $(x_{1:K}, u_{1:K})$. This posterior is a TMND where only the variables $u_{1:K}$ are restricted to the positive orthant. The TMND is approximated as a multivariate normal distribution whose parameters

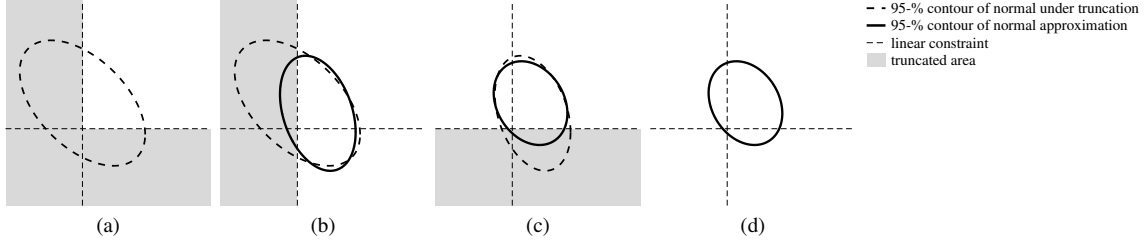


Fig. 3. The sequential truncation method for approximating a truncated normal distribution with a normal distribution: (a) the original normal distribution's contour ellipse that contains 95 % of the probability, and the truncated area in gray, (b) the first applied truncation in gray, and the 95-% contour of the resulting normal approximation, (c) the second applied truncation in gray, and the 95-% contour of the normal approximation, (d) the final normal approximation.

Table I
OPTIMAL SEQUENTIAL TRUNCATION TO THE POSITIVE ORTHANT

```

1: Inputs:  $\mu$ ,  $\Sigma$ , and the set of the truncated components' indices  $\mathcal{T}$ 
2: while  $\mathcal{T} \neq \emptyset$  do
3:    $k \leftarrow \operatorname{argmin}_i \{\mu_i / \sqrt{\Sigma_{ii}} \mid i \in \mathcal{T}\}$ 
4:    $\xi \leftarrow \mu_k / \sqrt{\Sigma_{kk}}$ 
5:   if  $\Phi(\xi)$  does not underflow to 0 then
6:      $\epsilon \leftarrow \phi(\xi) / \Phi(\xi)$   $\triangleright \phi$  is the PDF of  $\mathcal{N}(0, 1)$ ,  $\Phi$  its CDF
7:      $\mu \leftarrow \mu + (\epsilon / \sqrt{\Sigma_{kk}}) \cdot \Sigma_{:,k}$ 
8:      $\Sigma \leftarrow \Sigma - ((\xi\epsilon + \epsilon^2) / \Sigma_{kk}) \cdot \Sigma_{:,k} \Sigma_{:,k}$ 
9:   else
10:     $\mu \leftarrow \mu + (-\xi / \sqrt{\Sigma_{kk}}) \cdot \Sigma_{:,k}$   $\triangleright \lim_{\xi \rightarrow -\infty} (\epsilon + \xi) = 0$  [30]
11:     $\Sigma \leftarrow \Sigma - (1 / \Sigma_{kk}) \cdot \Sigma_{:,k} \Sigma_{:,k}$   $\triangleright \lim_{\xi \rightarrow -\infty} (\xi\epsilon + \epsilon^2) = 1$  [30]
12:   end if
13:    $\mathcal{T} \leftarrow \mathcal{T} \setminus \{k\}$ 
14: end while
15: Outputs:  $\mu$  and  $\Sigma$ ;  $(\mu, \Sigma) \leftarrow \operatorname{rec\_trunc}(\mu, \Sigma, \mathcal{T})$ 

```

are obtained using the sequential truncation. This approximation enables recursive forward filtering and the use of RTSS's backward smoothing step that gives normal approximations to the marginal smoothing posteriors $q_{xu}(x_k, u_k) \approx \mathcal{N}([x_k; u_k]; z_{k|K}, Z_{k|K})$. After the iterations converge, the variables $u_{1:K}$ are integrated out to get the approximate smoothing posteriors $\mathcal{N}(x_k; x_{k|K}, P_{k|K})$, where the parameters $x_{k|K}$ and $P_{k|K}$ are the output of the skew- t smoother (STS) algorithm in Table II. STS can be restricted to an online recursive algorithm to synthesize a filter which is summarized in Table III. In the filter, the output of a filtering step is also a TMND which in analogy to STS is approximated by a multivariate normal distribution to have a recursive algorithm. Using sequential truncation, the TMND $q_{xu}(x_k, u_k)$ is approximated by a normal distribution, and the parameters of the marginals $\mathcal{N}(x_k; x_{k|k}, P_{k|k})$ are the outputs of the skew- t filter (STF) algorithm in Table III.

IV. SIMULATIONS

Our numerical simulations use satellite navigation pseudo-range measurements of the model

$$[y_k]_i = \|s_i - [x_k]_{1:3}\| + [x_k]_4 + [e_k]_i, [e_k]_i \stackrel{\text{iid}}{\sim} \text{ST}(0, 1 \text{ m}, \delta \text{ m}, 4) \quad (16)$$

where $s_i \in \mathbb{R}^3$ is the i th satellite's position, $[x_k]_4 \in \mathbb{R}$ is bias with prior $\mathcal{N}(0, (0.75 \text{ m})^2)$, and $\delta \in \mathbb{R}$ is a parameter. The model is linearized using the first order Taylor polynomial approximation, and the linearization error is negligible because the satellites are far. The satellite constellation of the Global Positioning System from the first second of the year 2015 provided by the International GNSS Service [33] is used with 8 visible satellites. The root-mean-square error (RMSE) is computed for $[x_k]_{1:3}$. The computations are made with MATLAB.

Table II
SMOOTHING FOR SKEW- t MEASUREMENT NOISE

```

1: Inputs:  $A$ ,  $C$ ,  $Q$ ,  $R$ ,  $\Delta$ ,  $\nu$ ,  $x_{1|0}$ ,  $P_{1|0}$ , and  $y_{1:K}$ 
2:  $A_z \leftarrow \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$ ,  $C_z \leftarrow \begin{bmatrix} C & \Delta \end{bmatrix}$ 
   initialization
3:  $\Lambda_{k|K} \leftarrow I_{n_y}$  for  $k = 1 \dots K$ 
4: repeat
   update  $q_{xu}(x_{1:K}, u_{1:K})$  given  $q_\Lambda(\Lambda_{1:K})$ 
5:   for  $k = 1$  to  $K$  do
6:      $Z_{k|k-1} \leftarrow \operatorname{blockdiag}(P_{k|k-1}, \Lambda_{k|K}^{-1})$ 
7:      $K_z \leftarrow Z_{k|k-1} C_z^T (C P_{k|k-1} C^T + \Delta \Lambda_{k|K}^{-1} \Delta^T + \Lambda_{k|K}^{-1} R)^{-1}$ 
8:      $\tilde{z}_{k|k} \leftarrow \begin{bmatrix} x_{k|k-1} \\ 0 \end{bmatrix} + K_z (y_k - C x_{k|k-1})$ 
9:      $\tilde{Z}_{k|k} \leftarrow (I - K_z C_z) P_{k|k-1}$ 
10:     $[z_{k|k}, Z_{k|k}] \leftarrow \operatorname{rec\_trunc}(\tilde{z}_{k|k}, \tilde{Z}_{k|k}, \{n_x + 1 \dots n_x + n_y\})$ 
11:     $x_{k|k} \leftarrow [z_{k|k}]_{1:n_x}$ ,  $P_{k|k} \leftarrow [Z_{k|k}]_{1:n_x, 1:n_x}$ 
12:     $x_{k+1|k} \leftarrow A x_{k|k}$ 
13:     $P_{k+1|k} \leftarrow A P_{k|k} A^T + Q$ 
14:   end for
15:   for  $k = K - 1$  down to  $1$  do
16:      $G_k \leftarrow Z_{k|k} A_z Z_{k+1|k}^{-1}$ 
17:      $z_{k|K} \leftarrow z_{k|k} + G_k (z_{k+1|K} - A_z z_{k|k})$ 
18:      $Z_{k|K} \leftarrow Z_{k|k} + G_k (Z_{k+1|K} - Z_{k+1|k}) G_k^T$ 
19:      $x_{k|K} \leftarrow [z_{k|K}]_{1:n_x}$ ,  $P_{k|K} \leftarrow [Z_{k|K}]_{1:n_x, 1:n_x}$ 
20:      $u_{k|K} \leftarrow [z_{k|K}]_{n_x + (1:n_y)}$ ,  $U_{k|K} \leftarrow [Z_{k|K}]_{n_x + (1:n_y), n_x + (1:n_y)}$ 
21:   end for
   update  $q_\Lambda(\Lambda_{1:K})$  given  $q_{xu}(x_{1:K}, u_{1:K})$ 
22:   for  $k = 1$  to  $K$  do
23:      $\Psi_k \leftarrow (y_k - C_z z_{k|K})(y_k - C_z z_{k|K})^T R^{-1} + C_z Z_{k|K} C_z^T R^{-1}$ 
        $+ u_{k|K} u_{k|K}^T + U_{k|K}$ 
24:     for  $i = 1$  to  $n_y$  do  $[\Lambda_{k|K}]_{ii} \leftarrow \frac{\nu_i + 2}{\nu_i + 1 + [\Psi_k]_{ii}}$  end for
25:   end for
26: until converged
27: Outputs:  $x_{k|K}$  and  $P_{k|K}$  for  $k = 1 \dots K$ 

```

A. Computation of TMND statistics

In this subsection we study the computation of the moments of the untruncated components of a TMND. One state and one measurement vector per Monte Carlo replication are generated from the model (16) with $\nu = \infty$ degrees of freedom (corresponding to skew-normal likelihood), prior $x \sim \mathcal{N}(0, \operatorname{diag}((20 \text{ m})^2, (20 \text{ m})^2, (0.22 \text{ m})^2, (0.1 \text{ m})^2))$, and 10000 replications. The compared methods are sequential truncations with the optimal truncation order (TOPT) and with random order (TRAND), the variational Bayes (VB), and the analytical formulas of [27] using MATLAB function `mvncdf` (MVNCDF). In TRAND any of the non-optimal constraints is chosen randomly at each truncation. VB is an update of the skew t variational Bayes filter (STVBF) [20] where $\Lambda_k = I$ and the VB iteration is terminated when the position estimate changes less than 0.005 m or at the 1000th iteration. The reference solution for the expectation value is a bootstrap particle filter (PF) update with 50000 samples.

Table III
FILTERING FOR SKEW- t MEASUREMENT NOISE

```

1: Inputs:  $A, C, Q, R, \Delta, \nu, x_{1|0}, P_{1|0}$ , and  $y_{1:K}$ 
2:  $C_z \leftarrow [C \ \Delta]$ 
3: for  $k = 1$  to  $K$  do
  initialization
4:  $\Lambda_{k|k} \leftarrow I_{n_y}$ 
5: repeat
  update  $q_{xu}(x_k, u_k)$  given  $q_\Lambda(\Lambda_k)$ 
6:  $Z_{k|k-1} \leftarrow \text{blockdiag}(P_{k|k-1}, \Lambda_{k|k}^{-1})$ 
7:  $K_z \leftarrow Z_{k|k-1} C_z^T (C P_{k|k-1} C^T + \Delta \Lambda_{k|k}^{-1} \Delta^T + \Lambda_{k|k}^{-1} R)^{-1}$ 
8:  $\tilde{z}_{k|k} \leftarrow \begin{bmatrix} x_{k|k-1} \\ 0 \end{bmatrix} + K_z (y_k - C x_{k|k-1})$ 
9:  $\tilde{Z}_{k|k} \leftarrow (I - K_z C_z) P_{k|k-1}$ 
10:  $[z_{k|k}, Z_{k|k}] \leftarrow \text{rec\_trunc}(\tilde{z}_{k|k}, \tilde{Z}_{k|k}, \{n_x + 1 \dots n_x + n_y\})$ 
11:  $x_{k|k} \leftarrow [z_{k|k}]_{1:n_x}, P_{k|k} \leftarrow [Z_{k|k}]_{1:n_x, 1:n_x}$ 
12:  $u_{k|k} \leftarrow [z_{k|k}]_{n_x+(1:n_y)}, U_{k|k} \leftarrow [Z_{k|k}]_{n_x+(1:n_y), n_x+(1:n_y)}$ 
  update  $q_\Lambda(\Lambda_k)$  given  $q_{xu}(x_k, u_k)$ 
13:  $\Psi_k \leftarrow (y_k - C_z z_{k|k})(y_k - C_z z_{k|k})^T R^{-1} + C_z Z_{k|k} C_z^T R^{-1}$ 
   $+ u_{k|k} u_{k|k}^T + U_{k|k}$ 
14: for  $i = 1$  to  $n_y$  do  $[\Lambda_{k|k}]_{ii} \leftarrow \frac{\nu_i + 2}{\nu_i + [\Psi_k]_{ii}}$  end for
15: until converged
16:  $x_{k+1|k} \leftarrow A x_{k|k}$ 
17:  $P_{k+1|k} \leftarrow A P_{k|k} A^T + Q$ 
18: end for
19: Outputs:  $x_{k|k}$  and  $P_{k|k}$  for  $k = 1 \dots K$ 

```

Fig. 4 shows the distributions of the estimates' differences from the PF estimate. The errors are given per cent of the PF's estimation error. The box levels are 5%, 25%, 50%, 75%, and 95% quantiles and the asterisks show minimum and maximum values. TOPT outperforms TRAND in the cases with high skewness, which reflects the result of Lemma 1. MVNCDF is more accurate than TOPT in the cases with high skewness, but MVNCDF's computational load is roughly 40 000 times that of the TOPT. This justifies the use of the sequential truncation approximation.

The order of the truncations in the sequential truncation algorithm affects the performance only when there are clear differences in the amounts of probability mass that each truncation truncates. Fig. 5 presents an example, where $\delta = 20$, and each measurement noise realization e has been generated from the skew normal distribution and then modified by

$$e_j = \min\{\min\{e_{1:n_y}\}, 0\} - c\sqrt{1 + 20^2}, \quad (17)$$

where j is a random index, and c is a parameter. A large c generates one negative outlier to each measurement vector, which results in one truncation with significantly larger truncated probability mass than the rest of the truncations. Fig. 5 shows the difference of TRAND error from TOPT error; i.e. a positive difference means that TOPT is more accurate. The errors here refer to distance from the PF estimate. The figure shows that with large c TOPT is more accurate than TRAND. Thus, the effect of the truncation ordering on the accuracy of the sequential truncation approximation is more pronounced when there is one truncation that truncates much more than the rest. This justifies our greedy approach and the result of Lemma 1 for ordering the truncations.

The approximation of the posterior covariance matrix is tested by studying the normalized estimation error squared (NEES) values [34, Ch. 5.4.2]

$$\text{NEES} = (x_{k|k} - x_k)^T P_{k|k}^{-1} (x_{k|k} - x_k), \quad (18)$$

where $x_{k|k}$ and $P_{k|k}$ are the filter's output mean and covariance matrix, and x_k is the true state. The distributions of the filters'

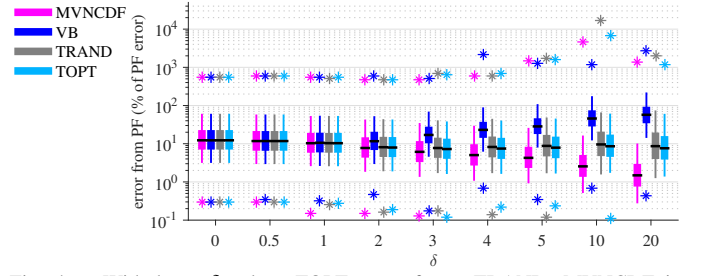


Fig. 4. With large δ values TOPT outperforms TRAND. MVNCDF is accurate but computationally heavy.

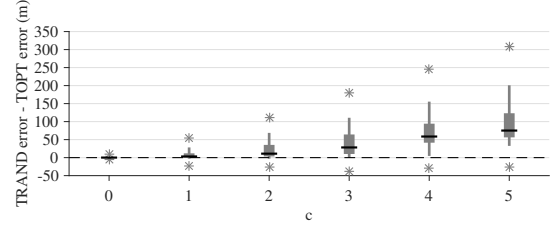


Fig. 5. TOPT outperforms TRAND when one negative outlier is added to the measurement noise vector because there is one truncation that truncates much more probability than the rest.

NEES values are shown by Fig. 6. If the covariance matrix is correct, the expected value of NEES is the dimensionality 3 [34, Ch. 5.4.2]. VB gets large NEES values when δ is large, which indicates that VB underestimates the covariance matrix. Apart from MVNCDF, TOPT and TRAND give NEES values closest to 3, so the sequential truncation provides a more accurate covariance matrix approximation than VB.

B. Skew- t inference

In this section, the proposed skew- t filter (STF) is compared with state-of-the-art filters using numerical simulations of a 100-step trajectory. The state model is a random walk with process noise covariance $Q = \text{diag}((qm)^2, (qm)^2, (0.2m)^2, 0)$, where q is a parameter. The compared methods are a bootstrap-type PF, STVBF [20], t variational Bayes filter (TVBF) [35], and Kalman filter (KF) with measurement validation gating [34, Ch. 5.7.2] that discards the measurement components whose normalized innovation squared is larger than the χ^2_1 -distribution's 99% quantile. The used KF parameters are the mean and variance of the used skew t -distribution, and the TVBF parameters are obtained by matching the degrees of freedom with that of the skew t -distribution and computing the maximum likelihood location and scale parameters for a set of pseudo-random numbers generated from the skew t -distribution. The results are based on 10 000 Monte Carlo replications.

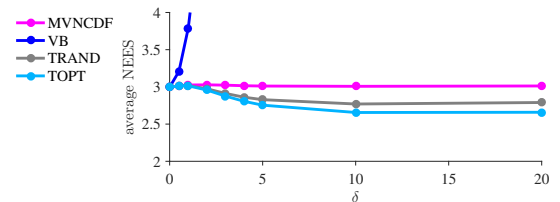


Fig. 6. TOPT's NEES is closer to the optimal value 3 than that of VB, so sequential truncation gives a more accurate posterior covariance matrix. $\delta = 0$ corresponds to the symmetric t -distribution hence, the comparison methods become identical.

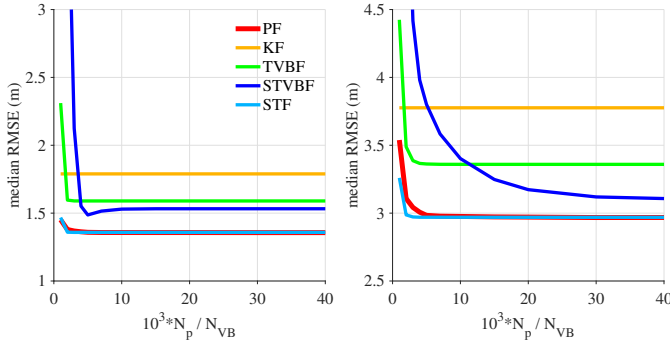


Fig. 7. STF converges in five iterations. The required number of PF particles can be more than 10 000. (left) $q=0.5$, $\delta=5$, (right) $q=5$, $\delta=5$.

Fig. 7 illustrates the filter iterations' convergence when the measurement noise components $[e_k]_i$ in (16) are generated independently from the univariate skew t -distribution. The figure shows that the proposed STF converges within five VB iterations and outperforms the other filters already with two VB iterations, except for PF that is the minimum-RMSE solution. Furthermore, Fig. 7 shows that STF's converged state is close to the PF's converged state in RMSE, and PF can require as many as 10 000 particles to outperform STF. STF also converges faster than STVBF when the process noise variance parameter q is large.

Fig. 8 shows the distributions of the RMSE differences from the STF's RMSE as percentages of the STF's RMSE. STF and TVBF use five VB iterations, and STVBF uses 30 VB iterations. STF clearly has the smallest RMSE when $\delta \geq 3$. Unlike STVBF, the new STF improves accuracy even with small q , which can be explained by the improved covariance matrix approximation.

The proposed smoother is also tested with measurements generated from (16). The compared smoothers are the proposed skew- t smoother (STS), skew- t variational Bayes smoother (STVBS) [20], t variational Bayes smoother (TVBS) [35], and the RTSS with 99% measurement validation gating [32]. Fig. 9 shows that STS has lower RMSE than the smoothers based on symmetric distributions. Furthermore, STS's VB iteration converges in five iterations or less, so it is faster than STVBS.

V. TESTS WITH REAL DATA

A. GNSS-based pseudorange positioning

Two GNSS (global navigation satellite system) positioning data sets were collected in central London (UK) to test the filters' performance in a challenging real-world satellite positioning scenario with numerous non-line-of-sight measurements. The data include time-of-arrival (TOA) based pseudorange measurements from GPS (Global Positioning System) satellites. Each set contains a trajectory that was collected by car using a u-blox M8 GNSS receiver. The lengths of the tracks are about 8 km and 10 km, the durations are about an hour for each, and measurements are received at about one-second intervals. The first track is used for fitting the filter parameters, while the second track is used for studying the filters' positioning errors. A ground truth was measured using an Applanix POS-LV220 system that improves the GNSS solution with tactical grade inertial measurement units. The GPS satellites' locations were obtained from the broadcast ephemerides provided by the International GNSS Service [33]. The algorithms are computed with MATLAB.

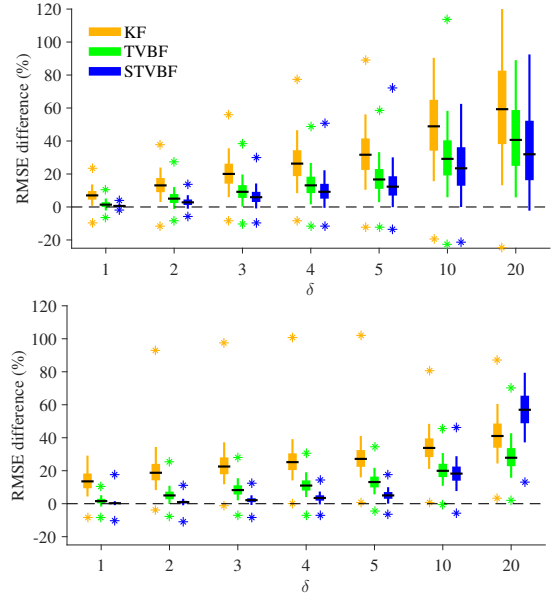


Fig. 8. STF outperforms the comparison methods with skew- t -distributed noise. RMSE differences from STF's RMSE per cent of the STF's RMSE. The differences increase as δ is increased. (upper) $q=0.5$, (lower) $q=5$.

The used state evolution model is the almost-constant velocity model for both the user position $l_k \in \mathbb{R}^3$ and the receiver clock error $b_k \in \mathbb{R}$ used in [36, Section IV]. Thus, the filter state is $x_k = [l_k^T \dot{l}_k^T b_k \dot{b}_k]^T \in \mathbb{R}^8$, and the state evolution model is

$$x_{k+1} = \begin{bmatrix} I_3 & D_k I_3 & O_{3 \times 2} \\ O_2 & I_3 & O_{3 \times 2} \\ O_{2 \times 3} & O_{2 \times 3} & \begin{bmatrix} 1 & D_k \\ 0 & 1 \end{bmatrix} \end{bmatrix} x_k + w_k, \quad (19)$$

where

$$w_k \sim \mathcal{N} \left(0, \begin{bmatrix} \frac{q^2 D_k^3}{3} I_3 & \frac{q^2 D_k^2}{2} I_3 & O_{3 \times 2} \\ \frac{q^2 D_k^2}{2} I_3 & q^2 D_k I_3 & O_{3 \times 2} \\ O_{2 \times 3} & O_{2 \times 3} & \begin{bmatrix} s_b D_k + \frac{s_f D_k^3}{3} & \frac{s_f D_k^2}{2} \\ \frac{s_f D_k^2}{2} & s_f D_k \end{bmatrix} \end{bmatrix} \right).$$

and D_k is the time difference of the measurements. The used parameter values are $q = 0.5$ m, $s_b = 70 \frac{\text{m}^2}{\text{s}}$, and $s_f = 0.6 \frac{\text{m}^2}{\text{s}^3}$. The initial prior is a normal distribution with mean given by the Gauss-Newton method with the first measurement and a large covariance matrix.

The measurement model is the same pseudorange model that is used in the simulations of Section IV, i.e.

$$[y_k]_i = \|s_{i,k} - [x_k]_{1:3}\| + [x_k]_7 + [e_k]_i, \quad (20)$$

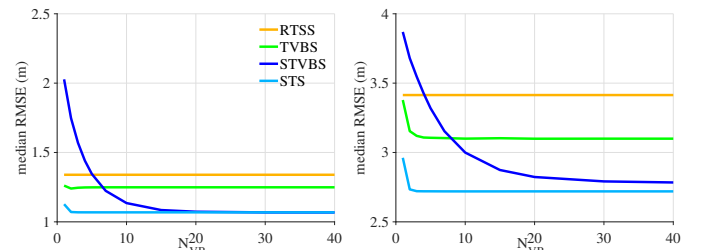


Fig. 9. Five STS iterations give the converged state's RMSE. (left) $q=0.5$, $\delta=5$, (right) $q=5$, $\delta=5$.

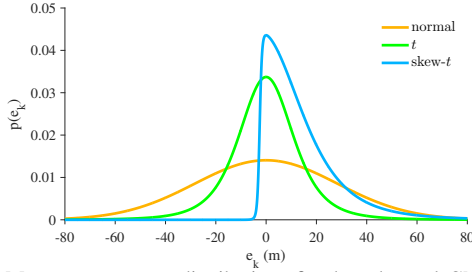


Fig. 10. Measurement error distributions fitted to the real GNSS data for normal, t , and skew- t error models. The modes are fixed to zero.

Table IV
FILTER PARAMETERS FOR REAL GNSS DATA

Skew- t , $\nu = 4$			t , $\nu_t = 4$		Normal	
μ (m)	σ (m)	δ (m)	μ_t (m)	σ_t (m)	μ_n (m)	σ_n (m)
-2.5	0.8	16.8	0	11.1	0	28.4

where $s_{i,k}$ is the 3-dimensional position of the i th satellite at the time of transmission. The measurement model is linearized with respect to x_k at each prior mean using the first order Taylor series approximation. The compared filters are based on three different models for the measurement noise e_k where

$$[e_k]_i \sim \text{ST}(\mu, \sigma^2, \delta, \nu); \quad (21)$$

$$[e_k]_i \sim \mathcal{T}(\mu_t, \sigma_t^2, \nu_t); \quad (22)$$

$$[e_k]_i \sim \mathcal{N}(\mu_n, \sigma_n^2). \quad (23)$$

The skew- t model (21) is the basis for STF and STVBF, the t model (22) is the basis for TVBF, and the normal model (23) is the basis for the extended KF (EKF) with 99% measurement validation gating. The pseudoranges are unbiased in the line-of-sight case, so the location parameters are fixed to $\mu_n = \mu_t = 0$. Furthermore, the degrees of freedom are fixed to $\nu = \nu_t = 4$ to compromise between outlier robustness and performance based on inlier measurements. The deviation parameter σ_n of the normal model was then fitted to the data using the expectation-maximization algorithm [37, Ch. 12.3.3] and the parameter σ_t of the t model as well as the parameters σ and δ of the skew- t model were fitted with the particle-Metropolis algorithm [37, Ch. 12.3.4]. The location parameter μ was obtained by numerically finding the point that sets the mode of the skew- t noise distribution to zero. These three error distributions' parameters are given in Table IV, and the PDFs are plotted in Fig. 10.

Fig. 11 shows the filter RMSEs as a function of the number of VB iterations. Both STF and TVBF converge within five VB iterations. The empirical CDF graphs of the user position errors with five VB iterations are shown in Fig. 12, and the RMSEs as well as the relative running times are given in Table V. The results show that modelling the skewness improves the positioning accuracy and is important especially for the accuracy in vertical direction. This can be explained by the sensitivity of the vertical direction to large measurement errors; due to bad measurement geometry the accuracy in the vertical direction is low even with line-of-sight measurements, so correct downweighting of erroneous altitude information requires careful modelling of the noise distribution's tails. The computational burden of our STF implementation with five VB iterations is about three times that of TVBF, but Fig. 11 shows that two STF iterations would already be enough to match TVBF's average RMSE.

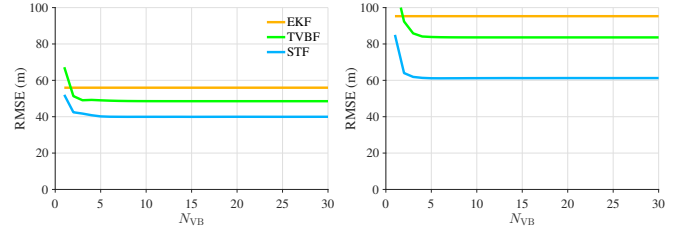


Fig. 11. RMSE of horizontal (left) and vertical (right) position for real GNSS data as a function of the number of VB iterations

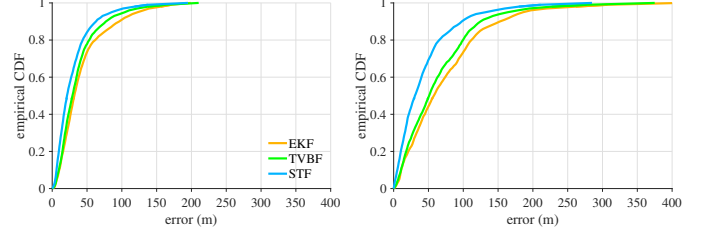


Fig. 12. Empirical error CDFs for the real GNSS data for the horizontal error (left) and the vertical error (right)

B. Ultra-wide band indoor positioning

Another application for STF is indoor localization using the TOA measurements of ultra-wide band (UWB) radios. We collected five test tracks in a laboratory environment with a high-precision optical reference positioning system¹ and three test tracks in a real university campus indoor environment. The measurement equipment is a Spoonphone smartphone [38] with Android 4.2 operating system and UWB channel 2 pulse radio (3993.6 MHz, 500 MHz bandwidth), and six BeSpoon UWB tags. The system uses two-way TOA ranging, thus no clock synchronization is required. The UWB measurement update and localization error computation is done with 2 Hz frequency. The algorithms are computed with MATLAB.

The novelty of this article compared to our previous article [8] is the STF algorithm. Here only TOA measurements are used, and the state evolution model is the random walk

$$x_{k+1} = x_k + w_k, \quad w_k \sim \mathcal{N}(0, \text{blockdiag}(q^2 \cdot I_2, q_a^2)), \quad (24)$$

where the state $x_k \in \mathbb{R}^3$ is the user position in east-north-up coordinates and $q = 0.5 \text{ ms}^{-\frac{1}{2}} \cdot \sqrt{0.5 \text{ s}}$ and $q_a = 1.4 \cdot 10^{-3} \text{ ms}^{-\frac{1}{2}} \cdot \sqrt{0.5 \text{ s}}$ are the process noise parameters. The initial position is assumed to be known.

Scaled by the speed of light, the two-way TOA gives a direct measurement of the distance between the UWB beacon and the user. Thus, the measurement model is

$$[y_k]_i = \|b_i - x_k\| + [e_k]_i, \quad (25)$$

where $y_k \in \mathbb{R}^{n_y}$ is the TOA-based distance vector, $b_i \in \mathbb{R}^3$ is the 3-dimensional position of the i th UWB beacon, and e_k is measurement noise. The measurement function is linearized at each prior mean.

We test the three alternative models (21)–(23) for the measurement noise. The estimation algorithm for the skew- t model (21) is the STF in Table III. The filter for the t -based model (22) is the TVBF [35] and for the normally distributed

¹High accuracy reference measurements are provided through the use of the Vicon real-time tracking system courtesy of the UAS Technologies Lab, Artificial Intelligence and Integrated Computer Systems Division (AIICS) at the Department of Computer and Information Science (IDA). <http://www.ida.liu.se/divisions/aiics/aiicssite/index.en.shtml>

Table V
THE RMSES AND RELATIVE RUNNING TIMES FOR REAL GNSS DATA

	EKF	TVBF	STF
RMSE _{horizontal} (m)	56	49	40
RMSE _{vertical} (m)	95	84	61
Running time	1	1.3	4.1

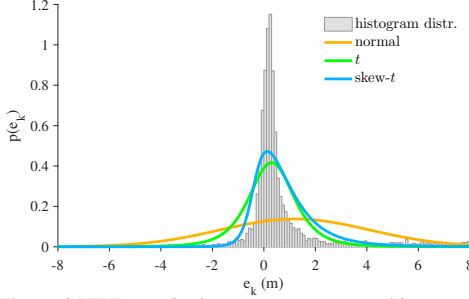


Fig. 13. The real UWB ranging's measurement error histogram distribution and the distributions fitted to the data for normal, t , and skew- t error models.

noise (23) the EKF with 99% measurement validation gating. The degrees of freedom parameters ν and ν_t are fixed to 4, and the parameters $\{\mu, \sigma, \delta, \mu_t, \sigma_t, \mu_n, \sigma_n\}$ are optimized to maximize the likelihood of the laboratory measurements. The maximum likelihood parameter values are given in Table VI, and the PDFs of the fitted distributions are compared with the error histogram in Fig. 13. The laboratory data are used for both parameter calibration and positioning tests to obtain a fair comparison of the optimal set-ups of each filtering algorithm. This eliminates the effect of possible differences in calibration and positioning data. To evaluate the performances at an independent data set, we also measured three test tracks in the campus of Tampere University of Technology with a rough reference position based on interpolation between timestamped locations.

Table VI
THE MAXIMUM LIKELIHOOD PARAMETER VALUES OF UWB POSITIONING BASED ON THE LABORATORY DATA SET

Skew- t , $\nu = 4$			t , $\nu_t = 4$		Normal	
μ (m)	σ (m)	δ (m)	μ_t (m)	σ_t (m)	μ_n (m)	σ_n (m)
-0.4	0.4	1.2	0.3	0.9	1.2	2.9

The VB-based filters also have the number of VB iterations N_{VB} as a parameter. Fig. 14 shows the filters' RMSEs averaged over all the data sets with different values of N_{VB} . Five iterations is sufficient for STF in UWB indoor positioning, but STF matches TVBF's accuracy already with two VB iterations. The RMSEs of the filters are given in Table VII. TVBF and STF use five VB iterations, and show significantly lower RMSE than EKF with measurement validation gating. Furthermore, the proposed STF has a lower RMSE than the TVBF. The campus track 3 was measured avoiding non-line-of-sight condition, so the difference between STF and TVBF is small at this track.

VI. EXTENSION TO MVST

The skew t -distribution has several multivariate versions. In [5]–[7] the PDF of the multivariate skew t -distribution (MVST) involves the CDF of a univariate t -distribution, while the definition of skew t -distribution given in [39] involves the CDF of a multivariate t -distribution. These versions of MVST are special cases of more general multivariate skew- t -type

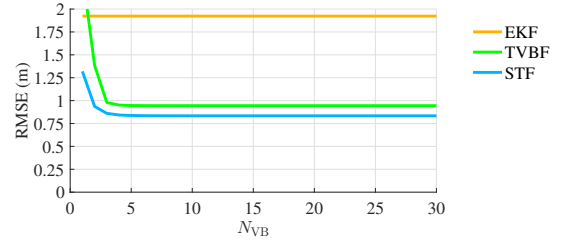


Fig. 14. Filter RMSEs for UWB indoor positioning as a function of the number of VB iterations

Table VII
THE AVERAGE RMSES IN METERS FOR UWB INDOOR POSITIONING

Route	EKF	TVBF	STF
Laboratory	2.12	0.92	0.79
Campus 1	1.79	1.18	1.06
Campus 2	1.24	0.80	0.72
Campus 3	1.75	1.00	0.98

distribution families, which include the multivariate canonical fundamental skew t -distribution (CFUST) [40] and the multivariate unified skew t -distribution [41]. A comprehensive review on the different variants of the MVST is given in [23].

The MVST variant used in this article is based on the CFUST discussed in [23], and it is the most general variant of the MVST. In this variant the parameter matrix $R \in \mathbb{R}^{n_z \times n_z}$ is a square positive-definite matrix, and $\Delta \in \mathbb{R}^{n_z \times n_z}$ is an arbitrary matrix. The PDF is

$$\text{MVST}(z; \mu, R, \Delta, \nu) = 2^{n_z} t(z; \mu, \Omega, \nu) T(\bar{z}; 0, L, \nu + n_z), \quad (26)$$

where $L = I_{n_z} - \Delta^T \Omega^{-1} \Delta$, $\Omega = R + \Delta \Delta^T$,

$$\begin{aligned} t(z; \mu, \Sigma, \nu) &= \frac{\Gamma\left(\frac{\nu + n_z}{2}\right)}{(\nu \pi)^{\frac{n_z}{2}} \det(\Sigma)^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{1}{\nu} (z - \mu)^T \Sigma^{-1} (z - \mu)\right)^{-\frac{\nu + n_z}{2}} \end{aligned} \quad (27)$$

is the PDF of the n_z -variate t -distribution and $T(z; \mu, \Sigma, \nu)$ its CDF, and

$$\bar{z} = \Delta^T \Omega^{-1} (z - \mu) \sqrt{\frac{\nu + n_z}{\nu + (z - \mu)^T \Omega^{-1} (z - \mu)}}. \quad (28)$$

The inference algorithms proposed in this paper can be extended to cover the case where the elements of the measurement noise vector are not statistically independent but jointly multivariate skew- t -distributed. When the measurement noise follows a MVST, i.e.

$$e_k \sim \text{MVST}(0, R, \Delta, \nu), \quad (29)$$

the filtering and smoothing algorithms presented in Tables III and II apply with slight modifications. At the core of this convenient extension is the fact that the MVST can be represented by a similar hierarchical model as in (6). However, the shape matrices Δ and R are not required to be diagonal, and the matrix Λ_k has the form $\lambda_k \cdot I_{n_y}$, where λ_k is a scalar with the prior

$$\lambda_k \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right). \quad (30)$$

Notice that when λ_k admits a small value, all the measurement components can potentially be outliers simultaneously unlike with the independent univariate skew- t components model. This difference is illustrated by the PDF contour plots in Fig. 15.

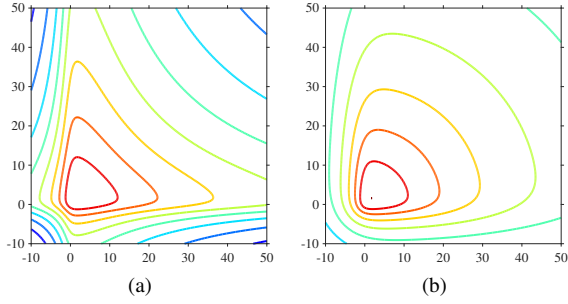


Fig. 15. PDF of bivariate measurement noise from (a) independent univariate skew- t components model (3) with $\Delta = 5I_2$, $R = I_2$, $\nu = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$ and (b) MVST model (29) with $\Delta = 5I_2$, $R = I_2$, $\nu = 4$.

The specific modification required by MVST measurement noise to the STS algorithm in Table II is replacing the line 24 by

$$\Lambda_{k|K} \leftarrow \frac{\nu + 2n_y}{\nu + \text{tr}\{\Psi_k\}} \cdot I_{n_y} \quad (31)$$

Similarly, the specific modification required by MVST measurement noise to the STF algorithm in Table III is replacing the line 14 by

$$\Lambda_{k|k} \leftarrow \frac{\nu + 2n_y}{\nu + \text{tr}\{\Psi_k\}} \cdot I_{n_y}. \quad (32)$$

VII. PERFORMANCE BOUND

A. Cramér–Rao lower bound

The Bayesian Cramér–Rao lower bound (CRLB) B is a lower bound for the mean-square-error (MSE) matrix of the state estimator \hat{x} of the random variable x using the observations y

$$\mathcal{M} = \mathbb{E}_{p(x,y)} [(x - \hat{x})(x - \hat{x})^T] \quad (33)$$

in the sense that the matrix difference $\mathcal{M} - B$ is positive semidefinite for any state estimator [42, Ch. 2.4]. The regularity conditions necessary for the positive-semidefiniteness to hold [42, Ch. 2.4] are the integrability of the first two partial derivatives of the joint distribution $p(x_{1:K}, y_{1:K})$ for an asymptotically unbiased estimator. These conditions are satisfied by the skew- t likelihood and the normal prior distribution, even though they do not hold for $p(x_{1:K}, u_{1:K}, \Lambda_{1:K}, y_{1:K})$ of the hierarchical model used in the proposed variational estimator due to restriction of $u_{1:K}$ to the positive orthant.

The filtering CRLB $B_{k|k}$ for the state-space model (1)–(2) follows the recursion [43]

$$B_{1|0} = P_{1|0} \quad (34a)$$

$$B_{k+1|k+1} = ((AB_{k|k}A^T + Q)^{-1} + \mathbb{E}_{p(x_k|y_{1:k-1})} [\mathcal{I}(x_k)])^{-1}, \quad (34b)$$

where $\mathcal{I}(e_k)$ is the Fisher information matrix of the measurement noise distribution. Furthermore, the smoothing CRLB for the state-space model (1)–(2) follows the recursion [43]

$$B_{k|K} = B_{k|k} + G_k(B_{k+1|K} - B_{k+1|k})G_k^T, \quad (35)$$

where

$$G_k = B_{k|k}A^TB_{k+1|k}^{-1}, \quad (36)$$

$$B_{k+1|k} = AB_{k|k}A^T + Q. \quad (37)$$

This coincides with the covariance matrix update of Rauch–Tung–Striebel smoother’s backward recursion [32].

The Fisher information matrix for the multivariate skew- t -distributed measurement noise of (29) is

$$\mathcal{I}(x) = C^T(R + \Delta\Delta^T)^{-\frac{T}{2}}E(R + \Delta\Delta^T)^{-\frac{1}{2}}C, \quad (38)$$

where

$$E = \mathbb{E}_{p(r)} \left[\frac{\nu + n_y}{\nu + r^T r} \left(I_{n_y} - \frac{2}{\nu + r^T r} r r^T + \tilde{R}_r \tilde{R}_r^T \right) \right] \quad (39)$$

with $r \sim \text{MVST}(0, I_{n_y} - \Theta\Theta^T, \Theta, \nu)$, $\Theta = (R + \Delta\Delta^T)^{-\frac{1}{2}}\Delta$, $A^{\frac{1}{2}}$ is a square-root matrix such that $A^{\frac{1}{2}}(A^{\frac{1}{2}})^T = A$, $A^{-\frac{1}{2}} \triangleq (A^{\frac{1}{2}})^{-1}$, $A^{-\frac{T}{2}} \triangleq ((A^{\frac{1}{2}})^{-1})^T$, and

$$\begin{aligned} \tilde{R}_r &= (T(\Theta^T r \sqrt{\frac{\nu + n_y}{\nu + r^T r}}; 0, I_{n_y} - \Theta^T \Theta, \nu + n_y))^{-1} \\ &\quad \times (I_{n_y} - \frac{1}{\nu + r^T r} r r^T) \Theta \\ &\quad \times \nabla_u T(u; 0, I_{n_y} - \Theta^T \Theta, \nu + n_y) \Big|_{u = \Theta^T r \sqrt{\frac{\nu + n_y}{\nu + r^T r}}}, \end{aligned} \quad (40)$$

where ∇_u is the gradient with respect to u . The derivation is given in Appendix B. The evaluation of the expectation in (39) is challenging with high-dimensional measurements due to the requirement to evaluate the CDF of the multivariate t -distribution and its partial derivatives. By the Woodbury matrix identity, the recursion (34) is equivalent to the covariance matrix update of the Kalman filter with the measurement noise covariance $(R + \Delta\Delta^T)^{\frac{1}{2}}E^{-1}((R + \Delta\Delta^T)^{\frac{1}{2}})^T$.

In the model (3) the measurement noise components are independently univariate skew- t -distributed. In this case the Fisher information is obtained by applying (38) to each conditionally independent measurement component and summing. The resulting formula matches with (38), the matrix E now being a diagonal matrix with the diagonal entries

$$\begin{aligned} E_{ii} &= \mathbb{E}_{p(r_i)} \left[\frac{\nu_i - r_i^2}{(\nu_i + r_i^2)^2} \right. \\ &\quad \left. + \frac{\theta_i^2}{1 - \theta_i^2} \frac{\nu_i^2}{(\nu_i + r_i^2)^3} \left(\tau_{\nu_i+1} \left(\frac{\theta_i}{\sqrt{1 - \theta_i^2}} r_i \sqrt{\frac{\nu_i + 1}{\nu_i + r_i^2}} \right) \right)^2 \right], \end{aligned} \quad (41)$$

where $r_i \sim \text{ST}(0, 1 - \theta_i^2, \theta_i, \nu_i)$ is a univariate skew- t -distributed random variable, $\theta_i = \Delta_{ii}/\sqrt{R_{ii} + \Delta_{ii}^2}$ and $\tau_\nu(x) = t(x; 0, 1, \nu)/T(x; 0, 1, \nu)$. Substituted into (38), this formula matches the Fisher information formula obtained for the univariate skew t -distribution in [44]. In this case only integrals with respect to one scalar variable are to be evaluated numerically.

B. Simulation

We study the CRLB in (34) of a linear state-space model with skew- t -distributed measurement noise by generating realizations of the model

$$x_{k+1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x_k + w_k, \quad w_k \sim \mathcal{N}(0, Q) \quad (42a)$$

$$y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} x_k + e_k, \quad e_k \sim \text{ST}(\mu, \sigma^2, \delta, \nu), \quad (42b)$$

where $x \in \mathbb{R}^2$ is the state, $Q = \begin{bmatrix} 1/3 & 1/2 \\ 1/2 & 1 \end{bmatrix}$ is the process noise covariance matrix, $y_k \in \mathbb{R}$ is the measurement, and ν

and δ_c are parameters that determine other parameters by the formulas

$$\mu = -\gamma\delta_c\sigma, \quad (43a)$$

$$\sigma^2 = \frac{\nu\omega^2}{\nu-2(1+\delta_c^2)-\gamma^2\delta_c^2}, \quad (43b)$$

$$\delta = \delta_c\sigma, \quad (43c)$$

$$\gamma = \sqrt{\frac{\nu}{\pi} \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}}. \quad (43d)$$

Thus, the measurement noise distribution is zero-mean and has the variance $\omega^2 = 5^2$. We generate 10000 realizations of a 50-step process, and compute the CRLB and mean-square-errors (MSE) of the bootstrap PF with 2000 particles and the STF. The CRLB and the MSEs were computed for the first component of the state at the last time instant $[x_{50}]_1$.

Fig. 16 shows the CRLB of the model (42). The figure shows that increase in the skewness as well as heavy-tailedness can decrease the CRLB significantly, which suggests that a nonlinear filter can be significantly better than the KF, which gives MSE 11.8 for all δ_c and ν . Fig. 17 shows the MSEs of PF and STF. As expected, when $\nu \rightarrow \infty$ and $\delta_c \rightarrow 0$, the PF's MSE approaches the CRLB. STF is only slightly worse than PF. The figures also show that although the CRLB becomes looser when the distribution becomes more skewed and/or heavy-tailed, it correctly indicates that modeling the skewness still improves the filtering performance.

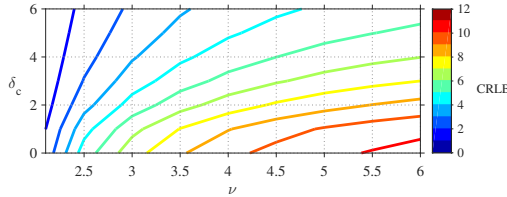


Fig. 16. The CRLB of the 50th time instant for the model (42) with a fixed measurement noise variance. Skewness and heavy-tailedness decreases the CRLB significantly.

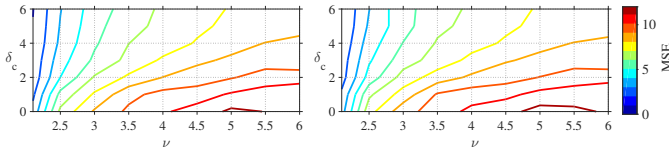


Fig. 17. The MSEs of PF (left) and STF's (right) are close to each other.

VIII. CONCLUSIONS

We have proposed a novel approximate filter and smoother for linear state-space models with heavy-tailed and skewed measurement noise distribution, and derived the Cramér-Rao lower bounds for the filtering and smoothing estimators. The algorithms are based on the variational Bayes approximation, where some posterior independence approximations are removed from the earlier versions of the algorithms to avoid significant underestimation of the posterior covariance matrix. Removal of independence approximations is enabled by the sequential truncation algorithm for approximating the mean and covariance matrix of truncated multivariate normal distribution. An optimal processing sequence is given for the sequential truncation. Simulations and real-data tests with GNSS outdoor positioning and UWB indoor positioning data show that the proposed algorithms outperform the state-of-the-art low-complexity methods.

REFERENCES

- [1] F. Gustafsson and F. Gunnarsson, "Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 41–53, July 2005.
- [2] B.-S. Chen, C.-Y. Yang, F.-K. Liao, and J.-F. Liao, "Mobile location estimator in a rough wireless environment using Extended Kalman-based IMM and data fusion," *IEEE Transactions on Vehicular Technology*, vol. 58, no. 3, pp. 1157–1169, March 2009.
- [3] M. Kok, J. D. Hol, and T. B. Schön, "Indoor positioning using ultra-wideband and inertial measurements," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, 2015.
- [4] K. Kaemarungsi and P. Krishnamurthy, "Analysis of WLAN's received signal strength indication for indoor location fingerprinting," *Pervasive and Mobile Computing*, vol. 8, no. 2, pp. 292–316, 2012, special Issue: Wide-Scale Vehicular Sensor Networks and Mobile Sensing.
- [5] M. D. Branco and D. K. Dey, "A general class of multivariate skew-elliptical distributions," *Journal of Multivariate Analysis*, vol. 79, no. 1, pp. 99–113, October 2001.
- [6] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t -distribution," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 65, no. 2, pp. 367–389, 2003.
- [7] A. K. Gupta, "Multivariate skew t -distribution," *Statistics*, vol. 37, no. 4, pp. 359–363, 2003.
- [8] H. Nurminen, T. Ardeschiri, R. Piché, and F. Gustafsson, "A NLOS-robust TOA positioning filter based on a skew- t measurement noise model," in *International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, October 2015, pp. 1–7.
- [9] S. Frühwirth-Schnatter and S. Pyne, "Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew- t distributions," *Biostatistics*, vol. 11, no. 2, pp. 317–336, 2010.
- [10] N. Counsell, M. Cortina-Borja, A. Lehtonen, and A. Stein, "Modelling psychiatric measures using skew-normal distributions," *European Psychiatry*, vol. 26, no. 2, pp. 112–114, 2010.
- [11] M. Eling, "Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models?" *Insurance: Mathematics and Economics*, vol. 51, no. 2, pp. 239–248, 2012.
- [12] Y. V. Marchenko, "Multivariate skew- t distributions in econometrics and environmental," Ph.D. dissertation, Texas A&M University, December 2010.
- [13] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, July 2000.
- [14] P. Naveau, M. G. Genton, and X. Shen, "A skewed Kalman filter," *Journal of Multivariate Analysis*, vol. 94, pp. 382–400, 2005.
- [15] H.-M. Kim, D. Ryu, B. K. Mallick, and M. G. Genton, "Mixtures of skewed Kalman filters," *Journal of Multivariate Analysis*, vol. 123, pp. 228–251, 2014.
- [16] J. Rezaie and J. Eidsvik, "Kalman filter variants in the closed skew normal setting," *Computational Statistics and Data Analysis*, vol. 75, pp. 1–14, 2014.
- [17] D. L. Alspach and H. W. Sorenson, "Nonlinear Bayesian estimation using Gaussian sum approximations," *IEEE Transactions on Automatic Control*, vol. 17, no. 4, pp. 439–448, Aug. 1972.
- [18] Y. Bar-Shalom and T. Fortmann, *Tracking and Data Association*, ser. Mathematics in Science and Engineering Series. Academic Press, 1988.
- [19] J. L. Williams and P. S. Maybeck, "Cost-function-based hypothesis control techniques for multiple hypothesis tracking," *Mathematical and Computer Modelling*, vol. 43, no. 9–10, pp. 976–989, May 2006.
- [20] H. Nurminen, T. Ardeschiri, R. Piche, and F. Gustafsson, "Robust inference for state-space models with skewed measurement noise," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1898–1902, November 2015.
- [21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2007.
- [22] M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Frühwirth, "Mean field variational Bayes for elaborate distributions," *Bayesian Analysis*, vol. 6, no. 4, pp. 847–900, 2011.
- [23] S. X. Lee and G. J. McLachlan, "Finite mixtures of canonical fundamental skew t -distributions – the unification of the restricted and unrestricted skew t -mixture models," *Statistics and Computing*, no. 26, pp. 573–589, 2016.
- [24] T. M. Cover and J. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.
- [25] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, Nov. 2008.
- [26] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.
- [27] G. Tallis, "The moment generating function of the truncated multi-normal distribution," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 23, no. 1, pp. 223–119, 1961.

- [28] A. Genz, "Numerical computation of rectangular bivariate and trivariate normal and t probabilities," *Statistics and Computing*, vol. 14, pp. 251–260, 2004.
- [29] A. Genz and F. Bretz, "Comparison of methods for the computation of multivariate t probabilities," *Journal of Computational and Graphical Statistics*, vol. 11, no. 4, pp. 950–971, 2002.
- [30] T. Perälä and S. Ali-Löytty, "Kalman-type positioning filters with floor plan information," in *6th International Conference on Advances in Mobile Computing and Multimedia (MoMM)*. New York, NY, USA: ACM, 2008, pp. 350–355.
- [31] D. J. Simon and D. L. Simon, "Constrained Kalman filtering via density function truncation for turbofan engine health estimation," *International Journal of Systems Science*, vol. 41, no. 2, pp. 159–171, 2010.
- [32] H. E. Rauch, C. T. Striebel, and F. Tung, "Maximum Likelihood Estimates of Linear Dynamic Systems," *Journal of the American Institute of Aeronautics and Astronautics*, vol. 3, no. 8, pp. 1445–1450, August 1965.
- [33] J. M. Dow, R. Neilan, and C. Rizos, "The international GNSS service in a changing landscape of global navigation satellite systems," *Journal of Geodesy*, vol. 83, no. 7, p. 689, February 2009.
- [34] Y. Bar-Shalom, R. X. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation, Theory Algorithms and Software*. John Wiley & Sons, 2001.
- [35] R. Piché, S. Särkkä, and J. Hartikainen, "Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate Student- t distribution," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2012.
- [36] P. Axelrad and R. Brown, "GPS navigation algorithms," in *Global Positioning System: Theory and Applications I*, B. W. Parkinson and J. J. Spilker Jr., Eds. Washington D.C.: AIAA, 1996, ch. 9.
- [37] S. Särkkä and J. Hartikainen, "On Gaussian optimal smoothing of nonlinear state space models," *IEEE Transactions on Automatic Control*, vol. 55, no. 8, pp. 1938–1941, August 2010.
- [38] The SpoonPhone. [Online]. Available: <http://spoonphone.com/en/>
- [39] S. K. Sahu, D. K. Dey, and M. D. Branco, "A new class of multivariate skew distributions with applications to Bayesian regression models," *Canadian Journal of Statistics*, vol. 31, no. 2, pp. 129–150, 2003.
- [40] R. B. Arellano-Valle and M. G. Genton, "On fundamental skew distributions," *Journal of Multivariate Analysis*, no. 96, pp. 93–116, 2005.
- [41] —, "Multivariate extended skew- t distributions and related families," *METRON - International Journal of Statistics*, vol. 68, no. 3, pp. 201–234, 2010.
- [42] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. New York: John Wiley & Sons, Inc., 1968.
- [43] M. Šimandl, J. Královic, and P. Tichavský, "Filtering, predictive, and smoothing Cramér–Rao bounds for discrete-time nonlinear dynamic systems," *Automatica*, vol. 37, pp. 1703–1716, 2001.
- [44] T. J. Di Ciccio and A. C. Monti, "Inferential aspects of the skew t -distribution," *Quaderni di Statistica*, vol. 13, pp. 1–21, 2011.
- [45] S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge, UK: Cambridge University Press, 2013.
- [46] R. Piché, "Cramér–Rao lower bound for linear filtering with t -distributed measurements," in *19th International Conference on Information Fusion (FUSION)*, July 2016, pp. 536–540.

APPENDIX A

DERIVATIONS FOR THE SKEW- t SMOOTHER

A. Derivations for q_{xu}

Eq. (9a) gives

$$\begin{aligned}
 \log q_{xu}(x_{1:K}, u_{1:K}) &= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) \\
 &+ \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
 &+ \sum_{k=1}^K \mathbb{E}_{q_\Lambda} [\log \mathcal{N}(y_k; Cx_k + \Delta u_k, \Lambda_k^{-1} R) \\
 &+ \log \mathcal{N}_+(u_k; 0, \Lambda_k^{-1})] + c \\
 &= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) + \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
 &- \frac{1}{2} \sum_{k=1}^K \mathbb{E}_{q_\Lambda} [(y_k - Cx_k - \Delta u_k)^T R^{-1} \Lambda_k (y_k - Cx_k - \Delta u_k) \\
 &+ u_k^T \Lambda_k u_k] + c
 \end{aligned} \tag{44}$$

$$\begin{aligned}
 &= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) + \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
 &- \frac{1}{2} \sum_{k=1}^K \{(y_k - Cx_k - \Delta u_k)^T R^{-1} \Lambda_{k|K} (y_k - Cx_k - \Delta u_k) \\
 &+ u_k^T \Lambda_{k|K} u_k\} + c
 \end{aligned} \tag{46}$$

$$\begin{aligned}
 &= \log \mathcal{N}(x_1; x_{1|0}, P_{1|0}) + \sum_{l=1}^{K-1} \log \mathcal{N}(x_{l+1}; Ax_l, Q) \\
 &+ \sum_{k=1}^K \{\log \mathcal{N}(y_k; Ax_k + \Delta u_k, \Lambda_{k|K}^{-1} R) \\
 &+ \log \mathcal{N}(u_k; 0, \Lambda_{k|K}^{-1})\} + c
 \end{aligned} \tag{47}$$

$$\begin{aligned}
 &= \log \mathcal{N} \left(\begin{bmatrix} x_1 \\ u_1 \end{bmatrix}; \begin{bmatrix} x_{1|0} \\ 0 \end{bmatrix}, \begin{bmatrix} P_{1|0} & 0 \\ 0 & \Lambda_{1|K}^{-1} \end{bmatrix} \right) \\
 &+ \sum_{l=1}^{K-1} \log \mathcal{N} \left(\begin{bmatrix} x_{l+1} \\ u_{l+1} \end{bmatrix}; \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_l \\ u_l \end{bmatrix}, \begin{bmatrix} Q & 0 \\ 0 & \Lambda_{l+1|K}^{-1} \end{bmatrix} \right) \\
 &+ \log \mathcal{N} \left(y_k; [C \quad \Delta] \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \Lambda_{k|K}^{-1} R \right) + c, u_{1:K} \geq 0,
 \end{aligned} \tag{48}$$

where $\Lambda_{k|K} \triangleq \mathbb{E}_{q_\Lambda} [\Lambda_k]$ is derived in Section A-B, and $u_{1:K} \geq 0$ means that all the components of all u_k are required to be nonnegative for each $k = 1 \cdots K$. Up to the truncation of the u components, $q_{xu}(x_{1:K}, u_{1:K})$ has thus the same form as the joint smoothing posterior of a linear state-space model with the state transition matrix $\tilde{A} \triangleq \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}$, process noise covariance matrix $\tilde{Q}_k \triangleq \begin{bmatrix} Q & 0 \\ 0 & \Lambda_{k+1|K}^{-1} \end{bmatrix}$, measurement model matrix $\tilde{C} \triangleq [C \quad \Delta]$, and measurement noise covariance matrix $\tilde{R} \triangleq \Lambda_{k|K}^{-1} R$. We denote the PDFs related to this state-space model with \tilde{p} .

It would be possible to compute the truncated multivariate normal posterior of the joint smoothing distribution $\tilde{p}(\begin{bmatrix} x_{1:K} \\ u_{1:K} \end{bmatrix} | y_{1:K})$, and account for the truncation of $u_{1:K}$ to the positive orthant using the sequential truncation. However, this would be impractical with large K due to the large dimensionality $K \times (n_x + n_y)$. A feasible solution is to approximate each filtering distribution in the Rauch–Tung–Striebel smoother’s (RTSS [32]) forward filtering step with a multivariate normal distribution by

$$\begin{aligned}
 \tilde{p}(x_k, u_k | y_{1:k}) &= \frac{1}{C} \mathcal{N} \left(\begin{bmatrix} x_k \\ u_k \end{bmatrix}; z'_{k|k}, Z'_{k|k} \right) \cdot \mathbb{I}[u_k \geq 0] \\
 &\approx \mathcal{N} \left(\begin{bmatrix} x_k \\ u_k \end{bmatrix}; z_{k|k}, Z_{k|k} \right)
 \end{aligned} \tag{49}$$

for each $k = 1 \cdots K$, where $\mathbb{I}[u_k \geq 0]$ is the Iverson bracket notation, C is the normalization factor, and $z_{k|k} \triangleq \mathbb{E}_{\tilde{p}}[\begin{bmatrix} x_k \\ u_k \end{bmatrix} | y_{1:k}]$ and $Z_{k|k} \triangleq \text{Var}_{\tilde{p}}[\begin{bmatrix} x_k \\ u_k \end{bmatrix} | y_{1:k}]$ are approximated using the sequential truncation. Given the multivariate normal approximations of the filtering posteriors $\tilde{p}(x_k, u_k | y_{1:k})$, by Lemma 2 the backward recursion of the RTSS gives multivariate normal approximations of the smoothing posteriors $\tilde{p}(x_k, u_k | y_{1:K})$. The quantities required in the derivations of Section A-B are the expectations of the smoother posteriors $x_{k|K} \triangleq \mathbb{E}_{q_{xu}}[x_k]$, $u_{k|K} \triangleq \mathbb{E}_{q_{xu}}[u_k]$, and the covariance matrices $Z_{k|K} \triangleq \text{Var}_{q_{xu}}[\begin{bmatrix} x_k \\ u_k \end{bmatrix}]$ and $U_{k|K} \triangleq \text{Var}_{q_{xu}}[u_k]$.

Lemma 2. Let $\{z_k\}_{k=1}^K$ be a linear-Gaussian process, and $\{y_k\}_{k=1}^K$ a measurement process such that

$$z_1 \sim \mathcal{N}(z_{1|0}, Z_{1|0}) \tag{51a}$$

$$\begin{aligned} z_k|z_{k-1} &\sim \mathcal{N}(Az_{k-1}, Q) \\ y_k|z_k &\sim (\text{a known distribution}). \end{aligned} \quad (51b) \quad (51c)$$

with the standard Markovianity assumptions. Then, if the filtering posterior $p(z_k|y_{1:k})$ is a multivariate normal distribution for each k , then for each $k < K$

$$z_k|y_{1:K} \sim \mathcal{N}(z_{k|K}, Z_{k|K}), \quad (52)$$

where

$$z_{k|K} = z_{k|k} + G_k(z_{k+1|K} - Az_{k|k}), \quad (53)$$

$$Z_{k|K} = Z_{k|k} + G_k(Z_{k+1|K} - AZ_{k|k}A^T - Q)G_k^T, \quad (54)$$

$$G_k = Z_{k|k}A^T(AZ_{k|k}A^T + Q)^{-1}, \quad (55)$$

and $z_{k|k}$ and $Z_{k|k}$ are the mean and covariance matrix of the filtering posterior $p(z_k|y_{1:k})$.

Proof: The proof is mostly similar to the proof of [45, Theorem 8.2]. First, assume that

$$z_{k+1}|y_{1:K} \sim \mathcal{N}(z_{k+1|K}, Z_{k+1|K}). \quad (56)$$

for some $k < K$. The joint conditional distribution of z_k and z_{k+1} is then

$$\begin{aligned} p(z_k, z_{k+1}|y_{1:k}) \\ = p(z_{k+1}|z_k) p(z_k|y_{1:k}) \quad || \text{Markovianity} \end{aligned} \quad (57)$$

$$= \mathcal{N}(z_{k+1}; Az_k, Q) \mathcal{N}(z_k; z_{k|k}, Z_{k|k}) \quad (58)$$

$$= \mathcal{N}\left(\begin{bmatrix} z_k \\ z_{k+1} \end{bmatrix}; \begin{bmatrix} z_{k|k} \\ Az_{k|k} \end{bmatrix}, \begin{bmatrix} Z_{k|k} & Z_{k|k}A^T \\ AZ_{k|k} & AZ_{k|k}A^T + Q \end{bmatrix}\right), \quad (59)$$

so by the conditioning rule of the multivariate normal distribution

$$\begin{aligned} p(z_k|z_{k+1}, y_{1:k}) \\ = \mathcal{N}(z_k; z_{k|k} + G_k(z_{k+1} - Az_{k|k}), \\ Z_{k|k} - Z_{k|k}A^T(AZ_{k|k}A^T + Q)^{-1}AZ_{k|k}) \end{aligned} \quad (60)$$

$$\begin{aligned} = \mathcal{N}(z_k; z_{k|k} + G_k(z_{k+1} - Az_{k|k}), \\ Z_{k|k} - G_k(AZ_{k|k}A^T + Q)G_k^T). \end{aligned} \quad (61)$$

We use this formula in

$$\begin{aligned} p(z_k, z_{k+1}|y_{1:K}) \\ = p(z_k|z_{k+1}, y_{1:K}) p(z_{k+1}|y_{1:K}) \end{aligned} \quad (62)$$

$$= p(z_k|z_{k+1}, y_{1:k}) p(z_{k+1}|y_{1:K}) \quad || \text{Markovianity} \quad (63)$$

$$\begin{aligned} = \mathcal{N}(z_k; z_{k|k} + G_k(z_{k+1} - Az_{k|k}), Z_{k|k} \\ - G_k(AZ_{k|k}A^T + Q)G_k^T) \mathcal{N}(z_{k+1}|z_{k+1|K}, Z_{k+1|K}) \end{aligned} \quad (64)$$

$$\begin{aligned} = \mathcal{N}\left(\begin{bmatrix} z_k \\ z_{k+1} \end{bmatrix}; \begin{bmatrix} z_{k|k} + G_k(z_{k+1|K} - Az_{k|k}) \\ \bullet \end{bmatrix}, \begin{bmatrix} Z_{k|k} + G_k(Z_{k+1|K} - AZ_{k|k}A^T - Q)G_k^T & \bullet \\ \bullet & Z_{k+1|K} \end{bmatrix}\right) \end{aligned} \quad (65)$$

$$\begin{aligned} = \mathcal{N}\left(\begin{bmatrix} z_k \\ z_{k+1} \end{bmatrix}; \begin{bmatrix} z_{k|k} + G_k(z_{k+1|K} - Az_{k|k}) \\ \bullet \end{bmatrix}, \begin{bmatrix} Z_{k|k} + G_k(Z_{k+1|K} - AZ_{k|k}A^T - Q)G_k^T & \bullet \\ \bullet & Z_{k+1|K} \end{bmatrix}\right), \end{aligned} \quad (66)$$

so

$$\begin{aligned} p(z_k|y_{1:K}) = \mathcal{N}(z_k; z_{k|k} + G_k(z_{k+1|K} - Az_{k|k}), Z_{k|K} \\ + G_k(Z_{k+1|K} - AZ_{k|k}A^T - Q)G_k^T) \end{aligned} \quad (67)$$

$$= \mathcal{N}(z_k; z_{k|K}, Z_{k|K}). \quad (68)$$

Because $z_K|y_{1:K} \sim \mathcal{N}(z_{K|K}, Z_{K|K})$, and because (56) implies (68), the statement holds by the induction argument. ■

B. Derivations for q_Λ

Eq. (9b) gives

$$\begin{aligned} \log q_\Lambda(\Lambda_{1:K}) = \sum_{k=1}^K \{ \mathbb{E}_{q_{xu}} [\log p(y_k|x_k, u_k, \Lambda_k) + \log p(u_k|\Lambda_k)] \\ + \log p(\Lambda_k) \} + c. \end{aligned} \quad (69)$$

Thus, $q_\Lambda(\Lambda_{1:K}) = \prod_{k=1}^K q_\Lambda(\Lambda_k)$.

In the model with independent univariate skew- t -distributed measurement noise components (3), the diagonal entries of Λ_k are separate random variables, as given in (6c). Therefore,

$$\begin{aligned} \log q_\Lambda(\Lambda_k) \\ = -\frac{1}{2} \mathbb{E}_{q_{xu}} [\text{tr}\{(y_k - Cx_k - \Delta u_k)(y_k - Cx_k - \Delta u_k)^T R^{-1} \Lambda_k\} \\ + \text{tr}\{u_k u_k^T \Lambda_k\}] + \sum_{i=1}^{n_y} \left(\frac{\nu_i}{2} \log[\Lambda_k]_{ii} - \frac{\nu_i}{2} [\Lambda_k]_{ii} \right) + c \end{aligned} \quad (70)$$

$$\begin{aligned} = -\frac{1}{2} \text{tr} \left\{ \left[((y_k - Cx_{k|K} - \Delta u_{k|K})(y_k - Cx_{k|K} - \Delta u_{k|K})^T \right. \right. \\ \left. \left. + [C \ \Delta] Z_{k|K} \begin{bmatrix} C^T \\ \Delta^T \end{bmatrix} \right) R^{-1} + (u_{k|K} u_{k|K}^T + U_{k|K}) \right] \Lambda_k \right\} \\ + \sum_{i=1}^{n_y} \left(\frac{\nu_i}{2} \log[\Lambda_k]_{ii} - \frac{\nu_i}{2} [\Lambda_k]_{ii} \right) + c \end{aligned} \quad (71)$$

$$= \sum_{i=1}^{n_y} \left(\frac{\nu_i}{2} \log[\Lambda_k]_{ii} - \frac{\nu_i + [\Psi_k]_{ii}}{2} [\Lambda_k]_{ii} \right) + c, \quad (72)$$

where

$$\begin{aligned} \Psi_k = (y_k - Cx_{k|K} - \Delta u_{k|K})(y_k - Cx_{k|K} - \Delta u_{k|K})^T R^{-1} \\ + [C \ \Delta] Z_{k|K} \begin{bmatrix} C^T \\ \Delta^T \end{bmatrix} R^{-1} + u_{k|K} u_{k|K}^T + U_{k|K}. \end{aligned} \quad (73)$$

Therefore,

$$q_\Lambda(\Lambda_k) = \prod_{i=1}^{n_y} \mathcal{G}\left([\Lambda_k]_{ii}; \frac{\nu_i}{2} + 1, \frac{\nu_i + [\Psi_k]_{ii}}{2}\right). \quad (74)$$

In the derivations of Section A-A, $\Lambda_{k|K} \triangleq \mathbb{E}_{q_\Lambda}[\Lambda_k]$ is required. $\Lambda_{k|K}$ is a diagonal matrix with the diagonal elements

$$[\Lambda_{k|K}]_{ii} = \frac{\nu_i + 2}{\nu_i + [\Psi_k]_{ii}}. \quad (75)$$

In the model (29) with multivariate skew- t -distributed measurement noise Λ_k is of the form $\lambda_k \cdot I_{n_y}$. There, λ_k is a scalar random variable, and there is just one degrees-of-freedom parameter ν , as given in (30). Therefore,

$$\begin{aligned} \log q_\Lambda(\lambda_k) \\ = -\frac{1}{2} \mathbb{E}_{q_{xu}} [\text{tr}\{(y_k - Cx_k - \Delta u_k)(y_k - Cx_k - \Delta u_k)^T R^{-1} \lambda_k\}] \\ - \frac{1}{2} \mathbb{E}_{q_{xu}} [\text{tr}\{u_k u_k^T \lambda_k\}] + \frac{\nu + 2n_y - 1}{2} \log \lambda_k - \frac{\nu}{2} \lambda_k + c \end{aligned} \quad (76)$$

$$= \frac{\nu + 2n_y - 1}{2} \log \lambda_k - \frac{\nu + \text{tr}\{\Psi_k\}}{2} \lambda_k, \quad (77)$$

where Ψ_k is given in (73). Thus,

$$q_\Lambda(\lambda_k) = \mathcal{G}\left(\lambda_k; \frac{\nu+2n_y}{2}, \frac{\nu+\text{tr}\{\Psi_k\}}{2}\right). \quad (78)$$

so the required expectation is

$$\Lambda_{k|K} = \frac{\nu+2n_y}{\nu+\text{tr}\{\Psi_k\}} \cdot I_{n_y}. \quad (79)$$

APPENDIX B

DERIVATION FOR THE FISHER INFORMATION OF MVST

Consider the multivariate skew- t measurement model $y|x \sim \text{MVST}(Cx, R, \Delta, \nu)$, where $C \in \mathbb{R}^{n_y \times n_x}$, $R \in \mathbb{R}^{n_y \times n_y}$, $\Delta \in \mathbb{R}^{n_y \times n_y}$, and $\nu \in \mathbb{R}^+$. The logarithm of the PDF of $y|x$ is

$$\log p(y|x) = \log(2^{n_y} / \det(\Omega)^{\frac{1}{2}}) + \log t(r; 0, I_{n_y}, \nu) \\ + \log T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y), \quad (80)$$

where $r = \Omega^{-\frac{1}{2}}(y - Cx)$ is a function of x and y , $\Omega = R + \Delta\Delta^T$, $L = I_{n_y} - \Delta^T \Omega^{-1} \Delta$, and $t(\cdot; \mu, \Sigma, \nu)$ and $T(\cdot; \mu, \Sigma, \nu)$ denote the PDF and CDF of the scaled non-central multivariate t -distribution with ν degrees of freedom. $A^{\frac{1}{2}}$ is a square-root matrix such that $A^{\frac{1}{2}}(A^{\frac{1}{2}})^T = A$, $A^{-\frac{1}{2}} \triangleq (A^{\frac{1}{2}})^{-1}$, and $A^{-\frac{T}{2}} \triangleq ((A^{\frac{1}{2}})^{-1})^T$.

The Hessian matrix of the term $\log t(r; 0, I_{n_y}, \nu)$ is derived in [46], and it is

$$\frac{d^2}{dx^2} \log t(r; 0, I_{n_y}, \nu) \\ = \frac{\nu+n_y}{\nu} C^T \Omega^{-\frac{T}{2}} \left(-\frac{1}{1+\frac{1}{\nu} r^T r} I_{n_y} + \frac{2/\nu}{(1+\frac{1}{\nu} r^T r)^2} r r^T \right) \Omega^{-\frac{1}{2}} C \quad (81)$$

$$= \frac{\nu+n_y}{\nu+r^T r} C^T \Omega^{-\frac{T}{2}} \left(-I_{n_y} + \frac{2}{\nu+r^T r} r r^T \right) \Omega^{-\frac{1}{2}} C \quad (82)$$

The term $\log T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y)$ can be differentiated twice using the chain rule

$$\frac{d^2 \log(f)}{dx^2} = \frac{1}{f} \frac{d^2 f}{dx^2} - \frac{1}{f^2} \left(\frac{df}{dx} \right)^T \frac{df}{dx}, \quad (83)$$

which gives

$$\frac{d^2}{dx^2} \log T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y) \\ = (T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-1} g(r) \\ - (T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-2} D_r^T P_r^T P_r D_r, \quad (84)$$

where the function g is antisymmetric because it is the second derivative of a function that is antisymmetric up to an additive constant,

$$P_r = \frac{d}{du} T(u; 0, L, \nu+n_y) \Big|_{u=\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}}, \quad (85)$$

and

$$D_r = \frac{d}{dx} \Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}} \quad (86)$$

$$= \sqrt{\frac{\nu+n_y}{\nu+r^T r}} \Delta^T \Omega^{-\frac{T}{2}} \left(\frac{1}{\nu+r^T r} r r^T - I_{n_y} \right) \Omega^{-\frac{1}{2}} C. \quad (87)$$

Because the function g is antisymmetric, $\int g(r)p(r) dy = 0$ for any symmetric function p for which the integral exists.

We now outline the proof of integrability of certain functions to show that the CRLB exists and fulfils the

regularity conditions given in [42, Ch. 2.4]. The integral $\int g(r) t(r; 0, 1, \nu) dy$ exists because the terms of g are products of positive powers of rational expressions where the denominator is of a higher degree than the nominator and derivatives of $T(u; 0, 1, \nu + n_y)$ evaluated at $\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}$, which is a bounded continuous function of y . The integral

$$\int (T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-1} D_r^T P_r^T P_r D_r \\ \times \frac{2}{\det(\Omega)^{\frac{1}{2}}} t(r; 0, I_{n_y}, \nu) dy$$

also exists because $(T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-1}$ and P_r are bounded and continuous and D_r is a positive power of a rational expression where the denominator is of a higher degree than the nominator. Similar arguments show the integrability of the first and second derivative of the likelihood $p(y|x)$, which guarantees that the regularity conditions of the CRLB are satisfied.

Thus, the expectation of (84) is

$$\mathbb{E}_{p(y|x)} \left[\frac{d^2}{dx^2} \log T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y) \right] \\ = \int g(r) \frac{2}{\det(\Omega)^{\frac{1}{2}}} t(r; 0, I_{n_y}, \nu) dy - \int \frac{2}{\det(\Omega)^{\frac{1}{2}}} t(r; 0, I_{n_y}, \nu) \\ \times (T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-1} D_r^T P_r^T P_r D_r dy \\ = \int 2g(r) t(r; 0, I_{n_y}, \nu) dr - \int 2t(r; 0, I_{n_y}, \nu) \\ \times (T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-1} D_r^T P_r^T P_r D_r dr \\ = - \mathbb{E}_{p(r|x)} \left[(T(\Theta^T r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-2} D_r^T P_r^T P_r D_r \right], \quad (88) \\ (89) \\ (90)$$

where $\Theta = \Omega^{-\frac{1}{2}} \Delta$, and $r|x \sim \text{MVST}(0, I_{n_y} - \Theta\Theta^T, \Theta, \nu)$ because $z \sim \text{MVST}(\mu, R, \Delta, \nu)$ implies $Az \sim \text{MVST}(A\mu, ARA^T, A\Delta, \nu)$. This gives

$$\mathbb{E}_{p(y|x)} \left[\frac{d^2}{dx^2} \log T(\Delta^T \Omega^{-\frac{T}{2}} r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y) \right] \\ = - C^T \Omega^{-\frac{T}{2}} \mathbb{E}_{p(r|x)} \left[\frac{\nu+n_y}{\nu+r^T r} \tilde{R}_r \tilde{R}_r^T \right] \Omega^{-\frac{1}{2}} C, \quad (91)$$

where

$$\tilde{R}_r = (T(\Theta^T r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}; 0, L, \nu+n_y))^{-1} (I_{n_y} - \frac{1}{\nu+r^T r} r r^T) \Theta \\ \times \left(\frac{d}{du} T(u; 0, L, \nu+n_y) \Big|_{u=\Theta^T r \sqrt{\frac{\nu+n_y}{\nu+r^T r}}} \right)^T, \quad (92)$$

where $L = I_{n_y} - \Theta\Theta^T$. Thus, the Fisher information for the measurement model $y|x \sim \text{MVST}(Cx, R, \Delta, \nu)$ is

$$\mathcal{I}(x) = \mathbb{E}_{p(y|x)} \left[-\frac{d^2}{dx^2} \log p(y|x) \right] \\ = C^T \Omega^{-\frac{T}{2}} \mathbb{E}_{p(r|x)} \left[\frac{\nu+n_y}{\nu+r^T r} (I_{n_y} - \frac{2}{(\nu+r^T r)^2} r r^T + \tilde{R}_r \tilde{R}_r^T) \right] \Omega^{-\frac{1}{2}} C, \quad (93) \\ (94)$$

where $r|x \sim \text{MVST}(0, I_{n_y} - \Theta\Theta^T, \Theta, \nu)$, $\Theta = \Omega^{-\frac{1}{2}} \Delta$, $\Omega = R + \Delta\Delta^T$, and \tilde{R}_r is defined in (92).